Supplementing Semantics with Statistics in the Personal Web

Mark H. Chignell and M. Ryan Kealey

Interactive Media Lab, Department of Mechanical and Industrial Engineering,

University of Toronto

1 Introduction

In her position paper for the First Symposium on the Personal Web", Ng (2010) describes the Personal Web as "user-sovereign web integration" resulting in the generation of dynamic and highly personalized web artifacts for visualizations and interactions. She envisions a process of "normalization and abstraction" that will converts entities on Web pages into integrate-able web elements suitable for Personal Web interactions. A "simple, well-defined meta-model" would then be used to generate visualizations and interactions based on the web integration results.

We will refer to Ng's vision of the personal web as "Plan A". Plan A assumes sufficiently well described Web content in the form of RDF annotation, and a sufficiently powerful semantics that can reason about, and link together, content from a variety of Websites. In this paper we propose a statistical approach as a kind of "Plan B" that may, in some situations, still provide useful personal web functionality in cases where Plan A fails.

2 Background

In his roadmap of the semantic web Berners-Lee (1998) characterized RDF as a necessary first step in regularizing semantics across websites. Ng (2010) posited that users could "gather personal linked data RDF graphs from several web domains and group them together into one semantic context". Berners-Lee saw an underlying (predicate) logic layer that would supplement RDF assertions with a powerful linking and querying capability, suggesting that different inference engines might be developed for different applications.

One view of the Personal Web is that it a personalized and contextualized version of the semantic web, motivated by the need to create smart interactions that are personalized according to the contexts and interests of individual users. Seen in this way, the Personal Web should be a more tractable version of the Semantic Web. The semantics of online shopping, travel, entertainment, self improvement and education, and personal health etc., are highly constrained relative to semantics in general. In this contextualized view, specialized inference engines make sense, perhaps forming a collection of expert systems (cf., Parsaye and Chignell, 1988) that provide the reasoning capabilities required by the Personal Web. However, while expert systems have had some success in specialized areas, they tend to be brittle, in that they fail to reason correctly when assumptions that they were built on don't hold and they can't fall back on the common sense knowledge that people tend to use when they don't have required expertise to deal with a problem.

The CYC project (Lenat and Guha, 1990) aimed at capturing "common sense knowledge" so that reasoning could be more generalized and less brittle. However, experience in artificial intelligence over recent decades suggests that it is much easier to develop reasoning engines than it is to represent the knowledge that is reasoned with. General-

ized reasoning of the type required for natural language understanding has yet to be achieved.

In summary, while the prospects for developing a semantic personal web are likely better than the semantic web as a whole, it is still a highly challenging task, since people and contexts vary along many different dimensions creating many different types of knowledge that need to be represented and reasoned with.

In natural language understanding, statistical reasoning has been used to supplement knowledge-based reasoning. With respect to processing of speech, Callison-Burch and Osborne (2003) claimed that: "Statistical techniques in speech recognition have so vastly outstripped the performance of their non-statistical counterparts that rule-based speech recognition systems are essentially no longer an area of research."

Sheth et al (2005) noted that exploiting heterogeneous data in the semantic web would require a broad range of semantics, which they classified into three forms: implicit; formal; powerful. In their characterization of "powerful" semantics: "Statistical techniques give us great insight into a corpus of documents or a large collection of data in general…. All derived relationships are statistical in nature and we only have an idea or a likelihood of their validity."

3 A Statistical Approach

In this paper we propose that explicit methods for representing semantics in the Personal Web through representation methods such as RDF and reasoning methods such as predicate logic be supplemented with statistical analysis (data analytics) and interactive visualization. The basic idea is that, in cases where explicit semantic reasoning cannot identify a sufficient set of integrated web elements that address the users need, interactive visualizations based on focused statistical analyses may provide users with "sufficiently convenient" overviews of relevant content and actions.

Appropriate information visualization has been proposed as a way to replace effortful thinking with simpler and more direct seeing (Card et al., 1999). A powerful demonstration of this idea, with direct manipulation of sliders to interactively manipulate the visualization (through inferred querying to the underlying database) was provided in the dynamic querying project (Williamson and Shneiderman, 1992). Since then there have been many techniques developed for interactive visualization (e.g., the elastic hierarchies developed by Zhao et al., 2005). New toolkits have been developed that greatly simplify the task of building interactive visualizations (e.g., Prefuse, Heer et al., 2005). These toolkits enable the development of applications that construct interactive visualizations to perform novel tasks such as tracking the popularity of baby names over time using census data (www.babynamewizard.com). Visualization toolkits are also leading to new forms of collaborative visualization, where data may be explored, interesting visualizations discovered, and then annotated and shared as part of larger discussions within blogs and communities of users (Heer et al, 2009).

The work of Casner (1991) and Wilkinson (2005) provide detailed perspectives on, and develop the idea of, task-based graphical presentation. In the proposed system, automatic analysis of the dataset would follow from analysis of the user's context and current tasks. Applications or 'task-based overlays' could then be developed on top of this system for facilitating the work of different different people and occupations, and the tasks associated with them. For instance, nurses in critical care units might use analytics-driven interactive visualization to keep better track of in-need patients (potentially reducing the incidence of "failure to rescue"). In a related example, emergency physicians could monitor the status of their other patients while they work with a particular patient. Continuing the healthcare example, but with a different occupational role, hospital administrators could monitor the overall situation and use general patterns of patient status to provide better measures of future bed requirements and availability, as well as identifying resource bottlenecks and the like.

4 A Healthcare Example

Healthcare is a domain where there is a huge amount of data, and where the available semantics and knowledge tend to apply to populations rather than individuals. Supplementing general medical knowledge with relevant views of how relevant peers of the current case respond to various treatments may provide improved clinical decision

support. The focus in this approach would be on the analytic and visualization tools that would take large volumes of (ideally stored in the form detailed electronic health records that include real-time data) healthcare data in the repository and make it available in the form of summaries and interactive visualizations. These tools would include parameters that could be set to allow healthcare professionals to customize their summarized view of the data. Algorithms could be built that would automatically generate visualizations of data depending on the context of the user and the dataset currently under review. Examples of automated statistical analyses (based on data patterns) that could generate appropriate visualizations were presented by Dan Rope at IBM University Day (Markham Ontario, April 2010). Making this approach work in realistic applications will likely involve extensive analyses of user requirements, followed by the prototyping and refinement of application user interfaces that address those requirements. The resulting application user interfaces then serve as specifications of what analytics and visualization capabilities are required in order to populate the user interface appropriately. We envisage the development process as a three legged stool where the work is supported by:

- Requirements analysis and user cases concerning the applications needed by healthcare professionals.
- Development of an analytics and visualization framework for automatically showing events, patterns and trends in large amounts of healthcare data.
- Prototyping and user testing of healthcare applications built using the analytics and visualization framework

The analytics and visualization framework will influence what can be prototyped, but at the same time knowledge of application requirements and user interface features may influence the functionality that is developed for the analytics and visualization framework. Similarly, application requirements will create a space of required functionality for visualization and analytics, but at the same time the analytics and visualizations that are developed may inspire and constrain application requirements.

For the analytics portion of the work we propose to utilize the SPSS statistical package, writing procedures that automate the use of particular techniques (such as time series analysis, cluster analysis and regression) to permit the conversion of the data into interpretable patterns, events, and trends. If possible we may build on the work that IBM/SPSS have already been doing in this area. In addition we propose to use Cognos tools to to organize data and to provide visual summaries of the key patterns and trends.

5 Conclusions

The Personal Web is a vision that requires identification and linking of integrateable Web entities. The success of Personal Web applications will depend to a large extent on the adequacy of these personalized and contextualized identification and linking operations. Characterizing Web semantics, and semantics in general, has proven to be a challenging problem and it remains to be seen how rapidly effective approaches to Personal Web Semantics will evolve. Personal Web statistics is presented as a complementary approach that will provide more flexibility in dealing with large, idiosyncratic, and poorly indexed datasets at the cost of requiring more interaction and exploration of the user. In many domains where users are highly motivated to achieve their tasks goals, this would seem to be a reasonable tradeoff.

Acknowledgements

The ideas expressed here have been influenced by a number of researchers ranging from Jock Mackinlay and others to Lars Grammel, Stephan Jou and Dan Rope. Much of our thinking on clinical decision support has grown out of earlier research with Sharon Straus. The research was supported by an IBM Faculty Award to the first author.

References

[1] Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. Scientific American, 284(5), 34-43.

[2] Callison-Burch, C. and Osborne, M. (2003). Statistical Natural Language Processing. In A. Farghaly (Ed.), Handbook for Language Engineers. CSLI Lecture Notes, University of Chicago Press.

[3] Card, S. K. , Mackinlay, J. D. and Shneiderman, B. (Eds., 1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc.

[4] Casner, S. M. 1991. Task-analytic approach to the automated design of graphic presentations. *ACM Trans. Graph*. 10, 2 (Apr. 1991), 111-151.

[5] Heer, J. (2005). Prefuse: a Toolkit for Interactive Information Visualization. In CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 421—430. NY: ACM Press.

[6] Heer, J., Viégas, F. B., and Wattenberg, M. 2009. Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Commun. ACM* 52, 1 (Jan. 2009), 87-97.

[7] Lenat, D. and Guha, R. V. (1990). Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley.

[8] Parsaye, K. and Chignell, M.H. "Expert Systems for Experts", New York, John Wiley & Sons, 1988.

[9] Sheth, A., Ramakrishnan, C., & Thomas, C. (2005), "Semantics for the Semantic Web: The implicit, the formal and the powerful," *International Journal on Semantic Web & Information Systems*, vol. 1, no. 1, pp. 1–18.

[10] Streitz, N. (1987). Cognitive compatibility as a central issue in human-computer interaction: Theoretical framework and empirical findings. In: G. Salvendy (Ed.), *Cognitive engineering in the design of human-computer interaction and expert systems*. Amsterdam: Elsevier Science. 75 - 82.

[11] Takeshita H, Davis D, Straus S. Clinical evidence at the point of care in acute medicine: a handheld usability case study. *Proceedings of the human factors and ergonomics society 46th annual meeting*, 2002, p. 1409-13.

[12] Wilkinson, L. (2005) *The Grammar of Graphics, 2nd Ed*. New York: Springer- Verlag.

[13] Williamson, C., and Shneiderman, B. (1992). The Dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system, Proc. ACM SIGIR'92 Conference, Copenhagen, Denmark, (June 1992), 338-346.

[14] Zhao, S., McGuffin, M.J., and Chignell, M.H. "Elastic Hierarchies: Combining Treemaps and Node-Link Diagrams," Information Visualization, IEEE Symposium on, p. 8, 2005 IEEE Symposium on Information Visualization (InfoVis 2005), 2005